

The REFOLD database: a tool for the optimization of protein expression and refolding

Michelle K. M. Chow¹, Abdullah A. Amin^{1,2}, Kate F. Fulton¹, Thushan Fernando⁴, Lawrence Kamau⁴, Chris Batty⁴, Michael Louca⁴, Storm Ho⁴, James C. Whisstock^{1,2,3}, Stephen P. Bottomley^{1,*} and Ashley M. Buckle^{1,2,*}

¹Department of Biochemistry and Molecular Biology, PO Box 13D, Monash University, Victoria 3800 Australia,

²Victorian Bioinformatics Consortium, PO Box 53, Monash University, Clayton, Victoria, 3800, Australia,

³ARC Centre for Structure and Functional Microbial Genomics, Monash University, Clayton, Victoria 3800, Australia and ⁴Swinburne University of Technology, Hawthorn Campus, Victoria, 3136, Australia

Received August 10, 2005; Revised and Accepted October 11, 2005

ABSTRACT

A large proportion of proteins expressed in *Escherichia coli* form inclusion bodies and thus require renaturation to attain a functional conformation for analysis. In this process, identifying and optimizing the refolding conditions and methodology is often rate limiting. In order to address this problem, we have developed REFOLD, a web-accessible relational database containing the published methods employed in the refolding of recombinant proteins. Currently, REFOLD contains >300 entries, which are heavily annotated such that the database can be searched via multiple parameters. We anticipate that REFOLD will continue to grow and eventually become a powerful tool for the optimization of protein renaturation. REFOLD is freely available at <http://refold.med.monash.edu.au>.

INTRODUCTION

In this post-genomic era, there is an imperative for the rapid production of recombinant native proteins for the advancement of structural biology and proteomics initiatives (1,2). Over-expression of recombinant proteins in bacteria is the routine method of choice. However, a considerable proportion of the proteins expressed in bacterial hosts aggregate to form *in vivo* inclusion bodies—this is particularly true for mammalian proteins (3). In such cases, inclusion bodies must be isolated, purified and solubilized with denaturing agents, followed by subsequent renaturation of the constituent protein. As such, there is a requirement for the development of highly

efficient methods of protein refolding, which minimize competing misfolding and aggregation reactions that lead to decreased yields. A wide range of protein refolding methodologies have been developed, utilizing simple dilution as well as more complex matrix-assisted methods and the addition of solutes to renaturing buffers (4–8).

The repertoire of refolding techniques is continually growing as medium- to high-throughput protein expression methods generate new proteins at an increasing rate. In order to exploit this wealth of data, a central repository for protein expression and refolding is critical. Traditionally, refolding experiments are conducted on an individual protein basis and are published in a non-standardized fashion. This presents a serious challenge for data retrieval and analysis. For example, simple searching for the details of a particular refolding protocol requires exhaustive manual inspection of the literature—data mining is practically impossible. In many cases, published data is essentially anecdotal. Thus the development of appropriate protocols for the expression and refolding of new proteins can be a relatively random and drawn-out process.

Structural genomics consortiums have exploited the wealth of refolding data emerging from high-throughput expression programs (9,10), but these data are not publicly accessible. Further, high-throughput structural genomics efforts can discard insoluble expressed proteins from their pipeline, as there is an abundance of soluble targets to focus on. However, in the broader research community, this position will not always be the case: many projects focus on targets that have important biological and medical significance, and the issue of solubility in bacterial expression cannot be SIDE-STEPPED. There are currently no publicly accessible repositories for protein refolding methodologies.

With the long-term goal of standardizing refolding data reporting and encouraging uniformity in the literature, we

*To whom correspondence should be addressed. Tel: +613 9905 3781; Email: ashley.buckle@med.monash.edu.au

Correspondence may also be addressed to Stephen P. Bottomley. Tel: +613 9905 4699; Email: steve.bottomley@med.monash.edu.au

A

Protein

Protein

Start With

- OR -

- none -

Class

Alpha

Family

- none -

Molecularity

Dimer

Species

Human

Disulphides

>=

1

Number of Record

=

Expression / Refolding

Expression Solubility

☐ Not Stated
☐ Soluble
☐ Partial
☐ Insoluble

Refolding Method

- none -

Redox Agent

- none -

pH

=

Temp (°C)

=

Protocol

Start With

Options

Use treeview

☐

Show commented entries

☐

Results per page

25

Show entries from

Last 3 months

New Search

Search

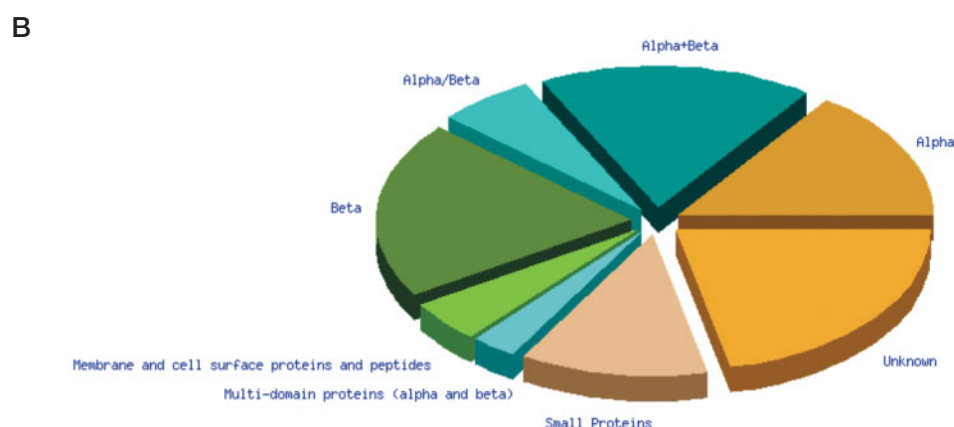













Figure 1. Web-based query interface in REFOLD. (A) Both simple and advanced querying can be performed. (B) Graphical representation of data. Portions of the chart can be selected to reveal the underlying data.

have created a freely available relational database, REFOLD (11,12), such that experimental refolding data can be rapidly deposited and readily accessed via a web-browser (<http://refold.med.monash.edu.au>). Entries have been extensively annotated with experimental details and protein structural characteristics such that the database can be searched via

multiple parameters. As a data repository, REFOLD allows new data to be rapidly deposited, validated and disseminated to the refolding and wider scientific community. We hope that by encouraging deposition, we can also make effective use of the vast amount of refolding data that goes unpublished because of the sheer difficulty in refolding many proteins.

A Search ResultsFound **304** results. Click on column header to sort. Click on  to expand or  to view refolding data.[Export to Excel](#)

Protein	Class	Organism	Family	Molecularity	Solubility	Refolding Method	Redox Agent
 1-aminocyclopropane-1-carboxylate synthase	Alpha/Beta	Cucurbita pepo (Squash)	GABA-aminotransferase-like	Dimer	insoluble	Dialysis	DTT
 3',5'-cyclic nucleotide phosphodiesterase	Alpha+Beta	Human	Cyclic nucleotide phosphodiesterase	Unknown	soluble	Dilution	DTT
 A-protein	Unknown	Aeromonas salmonicida	Unknown	Monomer	insoluble	Dialysis	None
 Acetylcholine Receptor alpha-Subunit	Membrane and cell surface proteins and peptides	Torpedo marmorata	Ligand-gated ionic channel	Monomer	insoluble	Dilution/Dialysis combination	GSH/GSSG
 Adrenomedullin	Unknown	Human	Unknown	Monomer	insoluble	Dilution	None
 Aequorin	Alpha	Aequorea victoria	Calmodulin-like	Monomer	insoluble	Column refolding: Nickel-chelating chromatography	Beta-mercaptoethanol
 Alkaline protease	Alpha/Beta	Aspergillus fumigatus	Subtilases	Unknown	insoluble	Dilution	None
 Alpha-1-Antichymotrypsin	Multi-domain proteins (alpha and beta)	Human	Serpins	Monomer	insoluble	Dilution/Dialysis combination	Beta-mercaptoethanol
 Alpha-1-Antichymotrypsin	Multi-domain proteins (alpha and beta)	Human	Serpins	Monomer	insoluble	Dilution	DTT

B Cluster/Browse by: [Class] [Family] [Organism] [Refolding Method]

Expand All Collapse All

Class (8)


- Alpha (15)
- Alpha+Beta (19)
- Alpha/Beta (13)
- Beta (21)
- Membrane and cell surface proteins and peptides (8)
- Multi-domain proteins (alpha and beta) (1)

Family

- Serpins (6)

Protein

- Alpha-1-Antichymotrypsin (2)
- Alpha-1-Antitrypsin (2)
- C1-Inhibitor (1)
- Neuroserpin (1)

Organism	Molecularity	MW (Da)	Solubility	Refolding Method	Redox Agent	pH	T(°C)
 Human	Monomer	44703.8	soluble	Dilution	None	7.8	4

- Pigment epithelium-derived factor (1)
- Plasminogen Activator Inhibitor-1 (1)

- Small Proteins (19)
- Unknown (3)

Figure 2. (A) Typical results of a search. Data can be sorted on any column and exported to MS Excel. (B) Search results can also be shown as a 'tree-view', which structures the data hierarchically according to structural class/family/protein.

The database also allows a level of data querying and access that is impossible using manual examination of the literature.

REFOLD DESCRIPTION AND USE

REFOLD was created using open-source MySQL relational database server software, version 4.1.10 (www.mysql.com), running on an Apple Dual 2.0 GHz G5/OS X Server (version 10.4.2). The database consists of 22 tables. A web-based query interface to the database was developed using the PHP5 programming language and PEAR database abstraction

classes, and is hosted on an Apple Dual 2.0 GHz G5/OS X Server (version 10.4.2) running Apache 1.3.33.

The website features a simple interface consisting of a central query box, a link to a more advanced search form (Figure 1A), a summary of the current database statistics, [including a section on recent updates, which is read directly from the REFOLD RSS (Really Simple Syndication) feed], and a small graphic depicting a breakdown of refold entries according to various criteria. This graphical representation of data is displayed automatically, but user-configurable graphical representation of data is also possible (Figure 1B).

PROTEIN	
Protein:	----- please select an option ----- * <input checked="" type="checkbox"/> Add New
SCOP Family:	Recombinase DNA-binding domain <input type="checkbox"/> Add New *
Protein Name:	HIV-1 Reverse transcriptase *
Protein Name (abbreviated):	HIV-1 RT *
Disulphides:	0 *
Molecularity:	Dimer *
Structure Solved?:	y *
MW (Da):	113694 *
UniProt ID: e.g., CD1A_HUMAN	Q6YA62_9HIV1 *

EXPRESSION	
Publication: (if unpublished select 'Unpublished')	Koller G, Graumann K, Kramer W, Sara M, Jungbauer A * <input type="checkbox"/> Add New
Project Aim:	Undefined *
Fusion:	None * <input type="checkbox"/> Add New
Enzymes:	
Expression Host:	E. coli *
Expression Strain:	JM105 *
Expression Level:	not stated *
Expression Temperature:	37 *
Expression Time:	4-5 h *
Expression Vector:	pKKRT66 *

REFOLDING	
Refolding Method:	Dilution * <input type="checkbox"/> Add New
Solubilization Buffer:	6M GdnHCl, 100mM TrisHCl pH 8.0, 100mM beta-mercaptoethanol, 1mM *
Wash Buffer:	n/a *
Refolding Buffer:	70mM Tris HCl pH 8.0, 20mM NaCl, 1mM EDTA, 1mM GdnHCl, 2mM redu *
Tag Cleaved?:	yes *
Protein Conc:	
Refolding pH:	8 *
Refolding Temperature:	4 *
Redox Agent:	GSH/GSSG * <input type="checkbox"/> Add New
Protocol:	E.coli cells were grown at 37degC in a 20L fermenter with a 14L net volume in M9ZB medium (Per L: 10g bactotryptone, 5g yeast extract, 5g NaCl, 1g NH4Cl,

Figure 3. Data deposition form, split into logical sections of protein (top), expression (middle) and refolding data (bottom). Only a portion of the form is shown here.

By default, a pie-chart shows a breakdown according to refolding technique—this provides immediate useful information to the user without performing a query. Most of the data can be visualized this way using pie or bar charts. Further, elements of the graph are hyperlinked directly to the data such that a mouse-click on a bar or pie segment will retrieve the data in the standard text format.

Data in REFOLD can be accessed by performing a simple search on any chosen term, or alternatively via a more specific advanced search. The database can be queried by numerous parameters, including gene species, refolding protocol and structural family (Figure 1A). For example, queries can be constructed which represent the following typical questions:

‘which proteins have been refolded using column-based methods?’; ‘show refolding records for proteins with a molecular mass greater than 100 kDa’; ‘find entries using chaperones’. The search results are shown in summary format and can be sorted on headings (Figure 2A). Detailed refolding data can be accessed by selecting individual rows.

Information on individual proteins relies heavily on the hierarchy used by the Structural Classification of Proteins database (SCOP) (13): proteins belong to families, which in turn belong to a structural class. The database can be browsed using the ‘Tree-view’ option in the search (Figure 1A): search results are then clustered according to a structural hierarchy (Figure 2B). Thus, the highly structured nature of REFOLD

may facilitate the initial scouting process for candidate refolding methodologies by allowing the inspection of protocols for proteins that have similar structural (e.g. domain content) or functional properties.

Search results can be exported in MS Excel format and REFOLD data is currently available as an RSS feed. This provides a simple mechanism whereby users can keep informed of recent updates to the database.

Open access to structured refolding data makes REFOLD a valuable resource for theoreticians looking for relationships between refolding success and protein characteristics such as charge and hydrophobicity. Data mining may uncover general predictive rules that could facilitate the refolding of novel proteins. It is important to bear in mind that any analysis using the relatively small set of proteins currently in the database is not yet statistically valid. At this stage, this is not the immediate aim of the database, and is better achieved using the significantly larger datasets generated by structural genomics programs (9,10) and proteome-scale purification efforts (3). However, in the longer term, as the number of entries in REFOLD grows, relationships and analyses arising from the constituent data will become more robust and informative.

DATA DEPOSITION

Registering a user account on REFOLD is fast, simple and secure, involving minimal data entry and taking no more than two minutes. Registered users can deposit their own refolding data using a 1-page form (Figure 3). The form is divided into three sections: Protein, Expression and Refolding. For each new protein entered into the database, various details of the biochemical and structural properties of the protein are recorded and for published refolding protocols, detailed reference to the relevant paper is also provided. Following this, experimental details of protein expression and refolding are entered, encompassing physical and chemical conditions, buffers used and efficiency of the methods employed. Depending on the format of data required, the form provides a mixture of text or number entry boxes and drop-down menus, often with the capacity to add new details if none of the existing options are applicable. Relevant links to other knowledge databases such as the UniProt (13), SCOP (14) and NCBI PubMed databases are also established through the data entry form. In addition to specified details, fields are provided for supplementary notes that may be useful to other users, and a full-text field is also included for comprehensive delineation of the expression and refolding protocols.

REFOLD also provides an extra opportunity for user input and dialogue by the inclusion of a comment box at the end of each refolding record. Thus researchers can provide feedback on existing data and perhaps offer further useful information, thereby creating a forum for scientific discourse and discussion.

CONCLUSIONS AND FUTURE DIRECTIONS

With the advent of the proteomic era and the subsequent myriad of functional and structural protein studies, the production of recombinant proteins is an essential tool

for the advancement of scientific and medical knowledge. Although the predictive use of correlations will prove useful for designing expression protocols, the majority of medium- to large-sized proteins expressed in bacteria will be insoluble and will ultimately require renaturation before analysis. Identifying the optimal refolding conditions and methodology then becomes the rate-limiting step in many studies, particularly if it relies on the use of oxidative and chaperone-assisted methods. REFOLD was created to assist in this stage, and we envisage that it will be of particular use to biologists working with proteins that are recalcitrant to renaturation. It is hoped that the usefulness of REFOLD encourages both deposition of refolding data in the same timeframe as publication, as well as continual deposition of unpublished material. In the long term, we aim to incorporate predictive functionality into REFOLD such that it will facilitate directly the design of refolding protocols. The following specific developments are in progress.

Sequence and domain annotation

We are currently modifying the database structure such that entries can be annotated by sequence and domain content. As the data grow, this step will be key to advancing REFOLD as a research tool.

Data mining tools

Although the advanced query tool allows searching across much of the database, we are developing a custom interface that will allow user-configurable queries against the whole dataset as well as user customization of how the results are displayed.

Interoperability with other databases

We are developing functionality that will integrate data from other sources, such as the comprehensive annotation of proteins in UniProt (13) supplied in XML format. It is important to note that a typical user of REFOLD is an experimentalist who most probably does not possess the skills (or indeed time) to navigate the vast array of protein sequence (e.g. pFam and UniProt) and structure (e.g. SCOP and PDB) information available. In order to facilitate this task, we will leverage data integration technologies set out in the eFamily project (<http://www.efamily.org.uk>). Accordingly, we will also make REFOLD data available in XML format for consumption by others. In the first instance, we are developing an XML schema that will provide a standard format for exchange.

Data visualization

As the dataset grows, visualization of text becomes cumbersome. This will require the development of more dynamic graphical representations of the data. In particular, graphical methods allowing the visualization of relationships between parameters, such as pH, pI, etc., will prove very useful.

Data deposition and validation

It is vital that new data is deposited in the same timeframe as publication. This means that the data becomes readily available to the community and amenable to analysis. Some validation logic is already built into the deposition process, providing both a useful service to the depositor as well as

an indication on data quality to users. We are working to streamline and automate this process.

The ongoing development of REFOLD with the clear aim of providing fast and user-friendly access to refolding data, integrated with a wide array of protein resources, will serve a worldwide community of life scientists faced with the demanding challenges of producing pure, active protein.

ACKNOWLEDGEMENTS

The authors would also like to acknowledge the contribution of all the researchers whose published data have been entered into REFOLD. This work was supported by grants from the National Health and Medical Research Council, the Victorian State Government and the Victorian Partnership for Advanced Computing. S.P.B. is a Monash University Senior Logan Fellow and NHMRC R.D. Wright Fellow. J.C.W. is a Monash University Logan Fellow and NHMRC Senior Research Fellow. K.F.F. is a NHMRC Peter Doherty Fellow. We thank Monash University, the National Health and Medical Research Council of Australia, the Australian Research Council, the Victorian State Government and the Victorian Partnership for Advanced Computing for support. Funding to pay the Open Access publication charges for this article was provided by the Australian Research Council.

Conflict of interest statement. None declared.

REFERENCES

1. Yee, A., Pardee, K., Christendat, D., Savchenko, A., Edwards, A.M. and Arrowsmith, C.H. (2003) Structural proteomics: toward high-throughput structural biology as a tool in functional genomics. *Acc. Chem. Res.*, **36**, 183–189.
2. Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Gerstein, M., Arrowsmith, C.H. and Edwards, A.M. (2000) Structural proteomics: prospects for high throughput sample preparation. *Prog. Biophys. Mol. Biol.*, **73**, 339–345.
3. Braun, P., Hu, Y., Shen, B., Halleck, A., Koundinya, M., Harlow, E. and LaBaer, J. (2002) Proteome-scale purification of human proteins from bacteria. *Proc. Natl Acad. Sci. USA*, **99**, 2654–2659.
4. Cabrita, L.D. and Bottomley, S.P. (2004) Protein expression and refolding—a practical guide to getting the most out of inclusion bodies. *Biotechnol. Annu. Rev.*, **10**, 31–50.
5. Mayer, M. and Buchner, J. (2004) Refolding of inclusion body proteins. *Methods Mol. Med.*, **94**, 239–254.
6. Middelberg, A.P. (2002) Preparative protein refolding. *Trends Biotechnol.*, **20**, 437–443.
7. Clark, E.D.B. (1998) Refolding of recombinant proteins. *Curr. Opin. Biotechnol.*, **9**, 157–163.
8. Clark, E.D. (2001) Protein refolding for industrial processes. *Curr. Opin. Biotechnol.*, **12**, 202–207.
9. Goh, C.S., Lan, N., Douglas, S.M., Wu, B., Echols, N., Smith, A., Milburn, D., Montelione, G.T., Zhao, H. and Gerstein, M. (2004) Mining the structural genomics pipeline: identification of protein properties that affect high-throughput experimental analysis. *J. Mol. Biol.*, **336**, 115–130.
10. Bertone, P., Kluger, Y., Lan, N., Zheng, D., Christendat, D., Yee, A., Edwards, A.M., Arrowsmith, C.H., Montelione, G.T. and Gerstein, M. (2001) SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. *Nucleic Acids Res.*, **29**, 2884–2898.
11. Buckle, A.M., Devlin, G.L., Jodun, R.A., Fulton, K.F., Faux, N., Whisstock, J.C. and Bottomley, S.P. (2005) The matrix refolded. *Nature Methods*, **2**, 3.
12. Chow, M.K., Amin, A.A., Fulton, K.F., Whisstock, J.C., Buckle, A.M. and Bottomley, S.P. (2005) REFOLD: an analytical database of protein refolding methods. *Protein Expr Purif.*, in press.
13. Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M. *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.
14. Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.